IJCAT International Journal of Computing and Technology, Volume 1, Issue 2, March 2014
ISSN : 2348 - 6090
www.IJCAT.org

A Technique for Offline Handwritten Character Recognition

[1] Shilpy Bansal, [2] Mamta Garg, [3] Munish Kumar

[1] Lecturer, Department of Computer Science Engineering, BMSCET, Muktsar, Punjab

[2]Assistant Professor, Department of Computer Science Engineering, BMSCET, Muktsar, Punjab

[3]Assistant Professor, Computer Science Department, P. U. Rural Centre, Kauni, Muktsar, Punjab

Abstract - Offline handwritten character recognition has been one of the most engrossing and challenging research areas in the field of pattern recognition in the recent years. Offline handwritten character recognition is a very problematic research area because writing styles may vary from one user to another. In this paper a proposed technique for offline handwritten Gurmukhi character recognition has been presented. The success rate of our proposed scheme depends upon the feature extraction technique which has been applied in this work. We have proposed a feature extraction technique named as Neighborhood Foreground Pixels Density technique. As there could be some insignificant feature values so to reject those we have used a dimensionality reduction technique namely Principal component analysis (PCA). Maximum recognition accuracy of 91.95% has been achieved with SVM (Radial Basis function kernel) classifier by using 10 fold cross validation test method.

***Keywords* - Classification, Feature extraction, SVM, MLP.**

1. Introduction

OCR stands for Optical character recognition. It is the process in which the handwritten, typewritten and printed text images are converted into machine editable text. It means it can convert any raster image of a document into a computer process able format such as ASCII format. The computer system which is having OCR system has various advantages like speed of input operations increases, data entry errors are reduced, storage space required by the paper documents are also reduced [1]. Optical character recognition can be classified based upon two major criteria: the data acquisition process (on-line or off-line) and the text type (machine-printed or handwritten) [2]. Offline character recognition system is that in which the handwritten character images are scanned from a surface (such as a sheet of paper) and stored into computer file in the form of two dimensional image representations which is called digital format and after that it is further processed to allow superior recognition [3]. Online character recognition system is that in which data is captured by the special device known as tablet digitizer which consists of electronic pen and an electrostatic sensitive writing surface and when this pen moves on the writing surface then these writing movements are converted into a sequence of signals and then send these signals to the connected computer [4].Offline character recognition system is different from Online character recognition system because Online contains more information about the writing style of a person such as speed, pressure, angle of a pen, direction and pressure against the system which is called order of the stroke. In handwritten character recognition, the characters are written by the individuals on a sheet of paper which is then stored into the computer by using scanner and after that these characters are recognized. Whereas In machine printed system the characters or text are typed from some input devices like any type of keyboard or typewriter etc. These texts are taken from some books, journals, newspapers etc. Handwritten character recognition is more complicated than machine printed because writing styles may vary from one user to another and large amount of noise will be occurs in the handwritten character recognition during the writing of the text and scanning of the document [5]. So Handwritten Character recognition has been one of the most engrossing and challenging research areas in the field of pattern recognition in the recent years. Even though, sufficient studies have been performed on handwritten character recognition system in few of the scripts like English, Arabic, Chinese and Indian scripts such as Devanagari, Bangla *etc.* but less work has been done on Gurmukhi script till date to the best of our knowledge. In this work, we have proposed recognition system for offline handwritten Gurmukhi characters.

2. Review Literature

Safi and Srinivagan [5] have proposed a new method to recognize offline Tamil handwritten characters. For feature extraction they proposed a zone based hybrid feature extraction technique which is the combination of number of horizontal, vertical, left diagonal and right diagonal lines, and total length of horizontal, vertical, left and right diagonal lines, number of intersection points and finally Euler number is also calculated. For classification Feed forward back propagation neural network is used. They conclude that they got better results than previous approaches. Tushar and Upadhyay [6]applied chain code based feature extraction method for recognition of handwritten cursive characters. After extracting the features these feature vectors are given to classifier named artificial neural network to classify the characters they got 80% recognition result. Aggarwal et al., [7] used novel methods of feature extraction for recognition of single isolated Devanagari character images. They used gradient feature extraction technique to extract the characters. SVM was used as a classifier in which they applied 5 fold validation schemes for classification and obtained good results. Kumar et al., [8] attempted to grade of different writers based on offline Gurmukhi characters written by them. They have used zoning, diagonal, directional, intersection, open end points and Zernike moment feature extraction techniques. For classification they used K-NN, HMM and Bayesian classifiers for grading of writers. The results obtained are in the form of average score achieved by the writers. Siddharth et al., [12] used three statistical features like zoning density, Projection histograms in which pixels are counted by moving in three directions like horizontal, vertical, left diagonal and right diagonal, distance profile features and Background Direction distribution features.

They also used different feature vectors by combining different features and then they applied SVM, K-NN and P-NN classifiers and used 5 fold cross validation for creating different testing sets. Impedovo [13] presented new advancements in zoning based recognition of handwritten characters. They describe various zoning topologies like static and dynamic. Zoning topologies can again be classified into slice based, shape based and hierarchal. They also discussed about zoning member functions like global and local. Global member function is further divided into order based and fuzzy. They also compared various zoning methods according to zoning topologies and zoning member functions. Kumar et al., [14] presents an efficient offline handwritten Gurmukhi character recognition system which is based on K-NN classifier. Diagonal and transition features are used to find the various features. Sharma et al., [15] proposed a Quadratic classifier based scheme for the recognition of off-line Devanagari handwritten characters. The features used in the classifier are obtained from the directional chain code information of the contour points of the characters. These chain code features are fed to the quadratic classifier for recognition. From the proposed scheme we obtained 80.36% recognition accuracy on Devanagari characters.

3. The Proposed Recognition System

Optical character recognition system includes various phases namely, Data acquisition, Digitization, Preprocessing, Feature extraction, and Classification. The block diagram of proposed system has been shown in Figure 1.

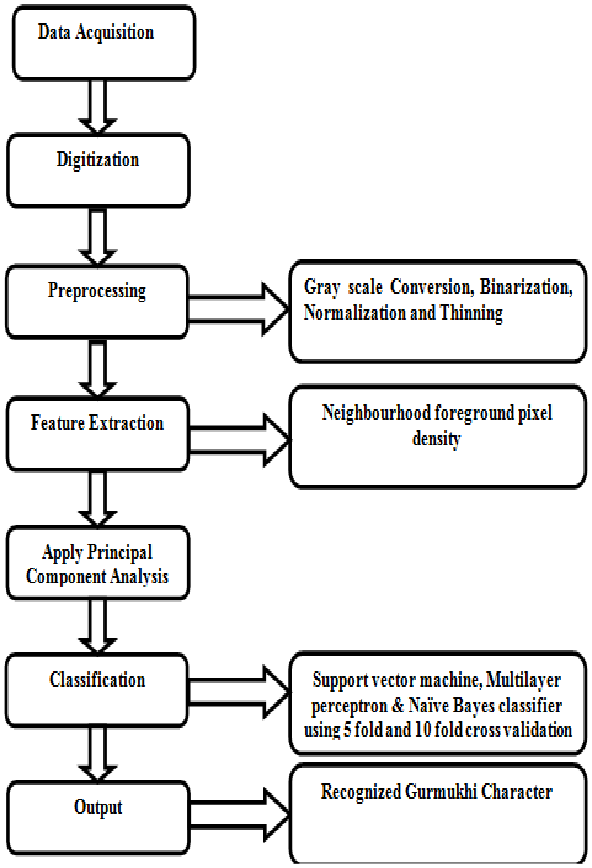
Data Acquisition
Digitization
Preprocessing
Gray scale Conversion, Binarization, Normalization and Thinning
Feature Extraction
Neighbourhood foreground pixel density
Apply Principal Component Analysis
Classification
Support vector machine, Multilayer perceptron & Naïve Bayes classifier using 5 fold and 10 fold cross validation
Output
Recognized Gurmukhi Character

Fig.1 Block diagram of offline handwritten Gurmukhi character recognition system

A. *Data Acquisition*: In our present work, we have developed handwritten database for recognition of Gurmukhi characters. We have taken the data set in which we have considered 35 basic characters of Gurmukhi alphabets. Samples of Gurmukhi characters were taken from 70 writers which were collected from people of different age groups having different professions.

B. *Digitization:* It is the process of converting the handwritten document into electronic form. This process is accomplished by using scanner or any other device after which it is converted into digital form.

C. *Preprocessing:* In preprocessing stage various techniques like gray scale, binarization, normalization and thinning are applied so that proper features are extracted.

D. *Feature Extraction:* Feature Extraction phase is the most
important phase in optical character recognition system as it is used for finding out various features of character which provides the relevant information about that character and it helps to recognize and classify that character with minimum amount of time and computation. After preprocessing phase, next the features are extracted by using the proposed feature extraction technique namely neighbourhood foreground pixel density.

Algorithm of Neighbourhood Foreground Pixel Density:

Recognition of offline handwritten Gurmukhi character.
Input: Pre-processed offline handwritten Gurmukhi images.

Output: Recognized Gurmukhi character.

1. Enter the Gray scale image and convert it into binary image.
2. Normalize the image into 100×100 size.
3. For each foreground pixel count the total no. of nearest foreground neighborhood pixels.
4. Update the Binary matrix with Number of neighboring pixels.
5. Divide this updated matrix into 100 zones with 10×10 size.
6. Then calculate the density value of foreground pixels with their updated values in each of the 100 zones. Then store this value as the feature element in this zone.
7. For the zones, that do not have foreground pixel, then consider the feature value as zero.
8. Normalize the feature values in scale from 0 to1.
9. These steps will give a feature vector of n elements.

A. *Principal Component Analysis*

Principal component analysis includes a mathematical procedure which transforms a number of correlated features into a number of uncorrelated features is called Principal Components [20]. PCA has wide applications in image processing, pattern recognition and computer vision. PCA is a technique which is used for dimensionality reduction. It means it eliminates the insignificant features and only significant features are left out for character recognition. So that these significant features show most of the variability of different characters. After applying PCA the number of features are minimized and one can get maximum infer from a less number of features and get maximum possible recognition rate in less time rather than a large number of features [21].

PCA is based on statistical method which uses covariance matrix, Eigen values and Eigen vectors. In this it extracts the R number of principal components having top R Eigen vectors with larger Eigen values. After forming feature vector of all the images we applied this dimensionality reduction technique named PCA through which out of 100 feature values only 78 feature values are formed.

B. *Classification*

This phase is the final phase of character recognition system. It is the process which to find a model to express and differentiate between classes to predict class labels of objects whose class label is unknown. This phase uses the feature vectors whose feature values are reduced by applying PCA and we have used three classifiers namely, Support Vector Machine, Naïve Bayes and Multilayer Perceptron to recognize handwritten Gurmukhi characters.

- **Support Vector Machine**

SVM classifier proves very good results in handwritten character recognition problems and this tool is also used for linear and non linear classification. Support vector machine have been extremely used by machine learning and pattern recognition communities.

The various fields in which this classifier is used are pattern recognition, face recognition and verification, speaker identification and verification, handwriting recognition, text categorization, prediction and image retrieval [22]. SVM is also called hyperplane classifier which is based on drawing the separating lines which separates the different objects that are having different class memberships [17]. In linear SVM the optimal decision hyperplane is that if the distance between nearest data point and the hyperplane is maximum, it means the generalization error of the classifier is minimized [23]. This is shown in the below given figure2.

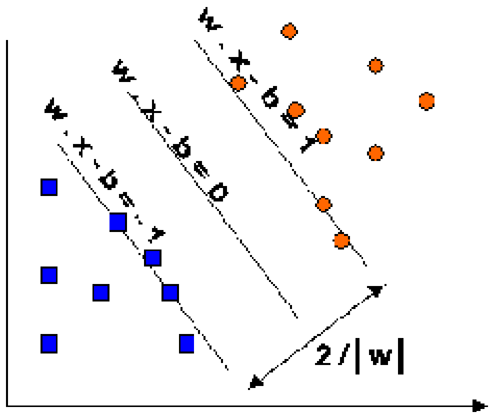

Fig. 2 Maximum margin Hyper planes for a SVM trained with samples from two classes [24]

SVM can be extended to the nonlinear classifier by mapping the data into high dimensional feature space by using various kernel functions. And there are four types of Kernel functions [24].

Linear Kernel: $K(x_i, x_j) = x_i^T x_j$
Polynomial Kernel: $\mathrm{K}(\mathrm{x_i}, \mathrm{x_j}) = (\mathrm{x_i^T x_j} + 1)^{\mathrm{d}}$
RBF Kernel:$K(x_i, x_j) = exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)$
Sigmoid Kernel:$K(x_i, x_j) = \tanh(\gamma x_i^T x_j - \theta)$

- **Multilayer Perceptron Classifier**

Neural networks have been intensively used in the area of pattern recognition and have increasingly received considerable attention in various areas such as signal processing, pattern recognition and automatic control [22]. Neural Network can be classified as feed forward and feed backward. And Multilayer perceptron classifier belongs to feed forward neural network. This classifier apply backpropagation algorithm which trained the classifier. Multilayer perceptron is a neural network that used a supervised learning technique [6]. This network has three types of layers: Input layer, Hidden layer and Output layer. Each layer is made up of collection of neurons. And each neuron is provided the input signals and with that the different weights are also assigned and then multiply the input values with their corresponding weights and finally summed them and then the sigmoid function which is a activation function is applied at each node [25]. Finally we get the output from the last layer which helps us to find the error and update our network according to the weight so that classification of unseen data can occur easily and accurately.

- **Naïve Bayes Classifier**

Naïve Bayes classifier is a type of Statistical classifier. This classifier is based on Bayes theorem. Bayes theorem calculates the maximum posterior hypothesis. This classifier calculates the probability of the character for a particular class and the character belongs to that class whose maximum posterior hypothesis is maximized.

4. Experimental Results and Discussion

In this section we have presented the results of offline handwritten Gurmukhi character recognition system. The results are generated by applying a proposed technique known as Neighbourhood Foreground pixels density technique. For this system we have collected 2450 samples of Gurmukhi character images. The 5-fold and 10-fold cross validation technique has been used for obtain the recognition accuracy.

Table.2 Recognition Accuracies of Various classifiers using 5 fold and 10 fold cross validations.

Classifier Used	5 fold Accuracy	10 fold Accuracy
SVM (Linear Kernel)	89.95%	90.81%
SVM (Poly. Kernel)	86.53%	87.51%
SVM (RBF Kernel)	91.79%	91.95%
SVM (Sigmoid Kernel)	82.24%	82.48%
MLP	86.32%	87.34%
Naïve Bayes	77.71%	77.63%

In 5-fold and 10-fold cross validation technique the data set is randomly divided into 5 and 10 mutually exclusive subsets of approximately equal size. We used one subset for testing and remaining subsets for training. Recognition results of proposed technique are shown in Table 2. The results show that we have achieved maximum accuracy of 91.95% using SVM classifier (RBF Kernel) with 10 fold cross validation technique

5. Conclusion and Future Scope

This paper proposes a new technique for offline handwritten Gurmukhi character recognition. For classification, we have used three different classifiers SVM, MLP and Naïve bayes, with 5 fold and 10 fold cross validations are used to recognize various characters and to achieve a better accuracy. SVM is also divided into 4 Kernels namely, Linear, Polynomial, RBF and Sigmoid Kernels. Using the neighbourhood foreground pixel density with PCA, we have achieved 91.95% accuracy with SVM (Polynomial Kernel) with 10 fold cross validation and with MLP we got 87.3% again with 10 fold cross validation and with naïve bayes we got 77.7% with 5 fold cross validation. This handwritten Gurmukhi character recognition work can also be extended to other Indian scripts Devanagri, Bengali, Tamil etc.

References

[1] Singh, S., Aggarwal, A. and Dhir, R. (2012), “Use of Gabor Filters for recognition of Handwritten Gurmukhi character”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, 5, pp. 234-240.

[2] Vaidya, S.A. and Bombade, B.R. (2013), “A Novel Approach of Handwritten Character Recognition using Positional Feature Extraction”, International Journal of Computer Science and Mobile Computing, Vol. 2, 6, pp. 179-186.

[3] Kannan, R.J. and Prabhakar, R. (2009), “A Comparative Study of Optical Character Recognition for Tamil Script”, European Journal of Scientific Research, Vol. 35, 4, pp. 570-582.

[4] Awasare, B., Patil, K. and Supriya, J. (2013), “Handwritten Script Recognition”, Proceeding of the International Conference on Emerging Trends in Engineering, pp. 30-33.

[5] Safi, L. A. and Srinivasagan, K. G. (2013), “Offline Tamil handwritten character recognition using zone based hybrid feature extraction technique”, International Journal of Computer Applications, Vol. 65, 1, pp. 12-16.

[6] Tushar, P. N. and Upadhyay, S. (2013), “Chain Code Based Handwritten Cursive Character Recognition System with Better Segmentation using Neural Network”,International Journal of Computational Engineering Research, Vol. 3, 5, pp. 60-63.

[7] Aggarwal, A., Rani, R. and Dhir, R. (2012), “Handwritten Devanagari character recognition using Gradient features”, International Journal of Advanced Researchin Computer Science and Software Engineering, Vol. 2, 5, pp. 85-90.

[8] Kumar, M., Jindal, M. K. and Sharma, R. K. (2011), “Classification of Characters and Grading Writers in Offline Handwritten Gurmukhi Script”, Proceeding of the International Conference on Image Information Processing, pp. 1-4.

[9] Sangame, Y. K., Ramteke, R. J. and Yogesh, V. G. (2011), “Recognition of isolated handwritten kannada characters using invariant moments and chain code”, World Journal of Science and Technology, Vol. 1, 8, pp. 115-120.

[10] Sharma, D. and Jhajj, P. (2010), “Recognition of Isolated Handwritten Characters in Gurmukhi Script”, International Journal of Computer Applications, Vol. 4, 8, pp. 9-17.

[11] Arora, S., Bhattacharjee, D., Nasipuri, M., Malik, L., Kundu, M. and Basu, D.K. (2010) “Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition”, Vol. 7, 6, pp. 18-26.

[12] Siddharth, K. S., Jangid, M., Dhir, R. and Rani, R. (2011), “Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features”, International Journal of Computer Science and Engineering, Vol. 3, 6, pp. 2332-2345.

[13] Impedovo, D., Pirlo, G. and Modugno, R. (2012), “New Advancements in Zoning-Based Recognition of Handwritten Characters”, Proceeding of the International Conference on Frontiers in Handwriting Recognition, pp. 661-665.

[14] Kumar, M., Jindal, M. K. and Sharma, R. K. (2011), “*k* - Nearest Neighbor Based Offline Handwritten Gurmukhi Character Recognition”, Proceeding of the International Conference on Image Information Processing, pp.1-4.

[15] Sharma, N., Pal, U., Kimura, F. and Pal, S. (2006), “Recognition of Off-Line Handwritten Devnagari Characters Using Quadratic Classifier”, Proceeding of the International Conference on Computer Vision Graphics and Image Processing, pp. 805-816.

[16] Khalsa, S. (2011), Punjabi Teaching [Online]. Available at http://www.sikhism.about.com/od/learntoreadgurmukhi/tp/gurmukhi_script.htm.

[17] Anjum, N., Bali, T. and Raj, B. (2013), “Design and simulation of handwritten multiscript character recognition”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, 7, pp. 2544-2549.

[18] Majumdar, A. and Chaudhuri, B. B. (2006), “A MLP Classifier for Both Printed and Handwritten Bangla Numeral Recognition”, Proceeding of the Indian Conference on Computer Vision, Graphics and Image Processing, 2006, pp. 796-804.

[19] Zhang, T. Y. and Suen, C. Y. (1984), “A Fast Parallel Algorithm for Thinning Digital Patterns”, Image Processing and Computer Vision, vol. 27, 3, 1984, pp. 236-239.

[20] Subbuthai, P., Periasamy, A. and Muruganand, S. (2012), “Identifying the Character by Applying PCA Method using Matlab”, International Journal of Computer Applications, Vol. 60, 1, pp. 8-11.

[21] Bahgat, S. F., Ghomiemy, S., Aljahdali, S. and Alotaibi, M. (2012), “A Proposed Hybrid Technique for Recognizing Arabic Characters”, International Journal of Advanced Research in Artificial Intelligence, Vol. 1, 4, pp. 35-43.

[22] Gazzah, S. and Amara, N. B. (2008) “Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script”, The International Arab Journal of Information Technology, vol. 1, 5, 2008, pp. 92-101.

[23] Shah, M. and Jethava, G. B. (2013), “A literature review on handwritten character recognition”, International Journal of Indian Streams Research Journal, Vol. 3, 2, pp. 1-19.

[24] Srivastava, D. K. and Bhambhu, L. (2009), “Data Classification Using Support Vector Machine”, International Journal of Theoretical and Advanced Information Technology, Vol. 12, 1, pp. 1-7.

[25] Sampath, A., Tripti, C. and Govindaru, V. (2012), “Freeman Code Based Online Handwritten Character Recognition for Malayalam using Backpropagation Neural Networks”, Advanced Computing: An International Journal, Vol. 3, 4, pp. 51-58.

[26] Das, N., Das, B., Sarkar, R., Basu, S., Kundu, M. and Nasipuri, M. (2010), “Handwritten BanglaBasic and Compound characterrecognition using MLP and SVM classifier”, Journal of Computing, Vol. 2, 2, pp. 109-115.